المحولات التوليدية العربية بنظرة تقنية رياضية بحثيّة مُختصرة

(مقدمة للنماذج اللغوية وأثرها وعلاقتها باللغة العربية)

عرض مختصر للنماذج رياضيات وحوسبيًّا وعلاقته بحثيًّا وفكريًّا

سري بن تيسير السباعي

مهندس باحث في معمل الروبوتات وإنترنت الأشياء في جامعة الأمير سلطان في السعودية

توطئة [جديدة]

مقدمة [القديمة]

الحمدُ لله وصلاة وسلاما على رسولِ الله أما بعدُ: فهذه ورقة معتصرة ألخِّص فيها أسسَ نماذجِ المحولاتِ المولِّدةِ فاكَّةِ التشفير[1] مُخصِّصًا منها ما كان متعلقا بالعربية. قسّمتها على ثلاث مقالات كلُّ مقالةٍ فصولٌ وكلُّ فصلٍ أبوابٌ وقبلهنَّ تقدمةٌ في ترجمة المصطلحات المُسْتعملة في الورقة.

أما تفصيل المقالات الثلاث:

فأولهن: في الكلام عن المدوّنات النصية العربية أنواعِها وطرقِ استنصاصها[2] وتجميعها ثم نتطرّق إلى طرق تمثيلها الرقمية للنموذج،

والثانية: نَصِفُ فيها التركيبَ الرياضيَّ وهيكلةَ النموذجِ المُحوّل فاكّ التشفير وأنواع بناءِ مكوّناتِه واختلافَها وألمح إلى مبحث سبب تعلُّم هذه النماذج، مختتمًا بذكرِ أسماء أشهرِ النَّماذج اللغوية العربية.

والثالثة في مَعْيَرة النماذج اللغوية الكبيرة واستعمالها على بيانات مخصوصةٍ والتطبيقات المُختلفة لذلك.

ويتخلل الفصول تنبيهات وفوائد حسب المناسب ان شاء الله.

واعلم أن هذه الورقة لم تُبيَّت فيها نية الاستقصاء ولا همة الجمع (فهذا يحتاج كتابا أو أكثر) وإنما غايتنا التنبيه والإلماح لعدد كبيرٍ من المصادر والمجالات والأفكار البحثية وربطها بالعربية نسأل الله العون وحسن الاختيار وصلى الله وسلم على نبيِّنا محمد وعلى آله وصحبه أجمعين.

وما من كاتبٍ إلا سيفنى ... ويُبقي الدهر ما كتبت يداه
فلا تكتب بكفِّك غير شيء ... يسرُّك في القيامة أن تراه[3]

---

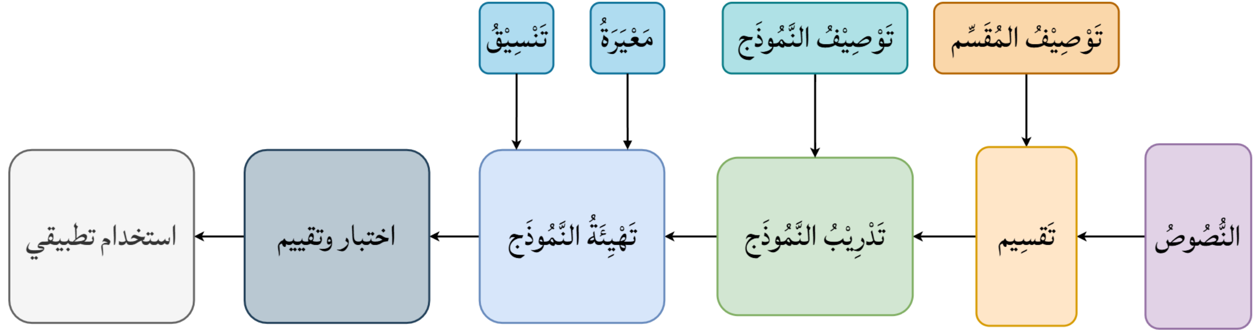
[1] واسمها Generative Transformer Decoder

[2] الاستنصاص من الفعل استنصص أي استفعل من نصَّ أي طلب النص

[3] يُنسب للشافعي ولعله لا يصح له ككثير مما نُسب له ونُسب أيضا لأمين الجندي الحمصي (ت 1256 هـ)

أفكار وإضافات جديدة

إضافة هندسة المدخلات واختراقاتها وكذا تقييم النماذج، وأيضًا مسألة استرجاع البيانات من النموذج المُدَرَّب، وزيادة المدونات النصية وذكر مسألة تقييم جودتها والدراسة الإحصائية (ان شاء الله) وطُرُق ذلك ما بين آلي ويدوي (إنسي)، والمقسمات مع العربية، التطبيقات العربية (انظر ورقة المرصد) . مبحث سلسلة الأفكار (الخواطر) لتحسين أجوبة النموذج CoT. وتضئيل (تقطير) النماذج Distillate. وأنظمة الراجّ.



فهرس الورقة (حدّث)

تقدمة في المصطلحات المترجمة

حرصا على استعمال اللسان العربي خالصا في هذه الورقة فإني قدمت هنا المصطلحات الأجنبية مع مقابلها العربي وسبب الترجمة أو التعريب إن وُجدَ ولعلي لم أستقصي كلَّ مُصطَلَحٍ لاشتهار بعضها عند الباحثين العرب ولألا أطيل الجدول فقد ذكرت ترجمة بعض المصطلحات وقتها.

المصطلح الإفرنجي (الأعجمي+الأجنبي)	الترجمة	السبب (التعليل)
Transformer	مُحَوِّلُ وجمعه محولات	هي الترجمة الحرفية للمصطلح المراد به نموذج شبكةِ عصبية يستخدم المنبهات.
Decoder Transformer	المحول فاكّ التشفير (الفاكّ اختصارا)	وهذه هيكلية النماذج المولدة عادة، وأشير هنا إلى أن الاسم الأشهر GPT أبدلناه بـ"متمم" وهي اختصار لـ"محوِّلٍ توليديّ مُدَرّب" وسيأتي
Large language model (LLM)	النموذج اللغوي الضخم (النلضة)	هي الترجمة الحرفية وقد نَحَتُّ كلمةَ "نلض" لاختصار الإستعمال في أثناء الكتابة وجمعها نلضات.
Byte pair Encoding (BPE)	التزويج البايتِيُّ (مُزَيبت)	استعضت عن ذكر التشفير أو الترميز في الاصطلاح اكتفاءً بوزن التفعيل في التزويج الدالِّ على عمله والله أعلم وهي خرزمةٌ لضغط النصوص.
Sentence piece	مُجمِل	الواقع أن هذه ليست خرزمةً مفردة وإنما إطارٌ مجموعٌ تحته أساليب وخِيارات متعددة لتدريب خرزمات تقسيم (تجزيء) النصوص.
Tokenization	تقسيم/تجزيء	تحويل النصوص المكتوبة لأجزاء (كُليمات) بحيث نعبّر عن الكل بمحدود.
Token	كُليمة	هي تصغير "كلمة" واقتديت فيها بمن عرب الفونيم بُ"صُوَيتة" [1] والكُليمة: الجزء النصي المحسوب والذي يُمثل مجموع كُليماته المدونة كاملةً بعدد محدود.
English	الكلزيّة	هذا استعمالي للغة الانجليزية أو الإنكليزي
Billion, Trillion …	جَلْفَ, دَلْفَ …	هذا تصميم عربي للكلمات الكلزية (مليار وغيرها) وتَعتمد على عدّ الجُمَّل: باستبدال أَلِفِ كلمةِ "أَلْف" بأحرف "أبجد هوّز" حيْثُ أن ترتيب الحروف هو $10^{ع\times3}$ فالعين مكان الحرف. مثلًا الدال هو الرابع فتصير $10^{4\times3} = 10^{12}$ وهكذا تِبَاعًا. [2]

Wikipedia	موسوكة	وهي موسوعة الشبكة العامة التي تحوي مقالات تغطي غالب المعارف الإنسانية، وهذه الكلمة مصكوكة ليَسهُلَ ذكرها.
internet	الشابكة	وهي الترجمة المعتمدة
OCR optical character recognition	التعرف البصري على المحارف	
Algorithm	خرزمة	استللنا من المصطلح المشهور بالخوارزميّات مادّة (خ-ر-ز-م) لفعل رباعي هو خرزم لتُستعملَ الكلمة على مُقتضيات التصريف العربي كُلٌّ بحسب سياقه وحاجته.
merges	دمجات	عدد يحدد عدد كُليمات المقسّم.
Finetune	معيرة أو دوزنة	هذه الترجمة فيها جوهر هذه العملية إذ مقصودها ترويض النموذج على بيانات أو أداء (سلوك) محدد. [وقد اقترحت أيضا توليف وتنغيم]
Instruct	مُهَيكَل	في سياق تطوير النماذج اللغوية فذا المصطلح يعني عرض النصوص بطريقة منهجية بِبُنية مُنمَّطة.
RLHF reinforcement learning from human feedback	التعلم المعزز من الإفادة البشرية (اختصارا تفاود) أو قُل من الإثراء البشري[4]	التفاود هو التفاعل وهذا الوزنُ صرفًا يعني المشاركة بين اثنين أو أكثر وهنا يتفاعل البشر مع النموذج حتى يجيب كما وعلى ما يحبون.
GPT generative pretrained transformer	محول توليدي مُدَرَّب (اختصارا متمم)	هذه الترجمة صدفَ أن أول حرفٍ منها نحتًا يُكوِّن كلمة ذات معنى ذاتي قريبٍ جدا من المعنى الأصلي فإن النموذج يُكمل أو يتمم الكلام المبتدأ (يوَلِّدُ).
Tensor	مرصوصة	تعميم لمفهوم المصفوفات تزيد أبعاده تبدأ من واحد (وهذا يُمثل المتجه) واثنين (المصفوفة) وأكثر وهو مشهور باسم تنسر [3]
Dot Product	الجُداء الداخلي	عملية بين مرصوصتين مشهورة في الجبر الخطي يُضرب فيها كل صف من المرصوصة الأولى بكل عمود في الثانية وتُجمع النتيجة. والجُداء أفصح من الضرب.
Neural network	شبكة عصبية (نحتًا شعصبونة)	سلسلة طبقات جُدائية تُتعَلَّمُ فيها أوزان تُقَرِّب معادلة معيَّنة صعبة الصياغة. تحاكي في عملها

[4] مستفاد من كتاب "المعالجة الآلية للنصوص العربية" الفصل الأول ص30.

		ظاهر عملِ العصبونات الحيوية
Parallelism	التآزي	أمور تتم بوقت واحد والأفصح بالهمزة تآزي بدلا من توازي كما في مختار الصحاح. [4]
Logits	مُحُض أو صُرُف	هذه الترجمة من ابتكاري وأول من وجدت ترجم كلمة logits الأستاذ طعيمة في كتابه معجم الذكاء الاصطناعي [5] بالسجلات وأنا أرى والله أعلم أن الذي ذكرته أقرب لأنَّ هذا المتجه يُحَوَّل إلى توزيع احتمالي فقد كان مَحضًا ثم شُكِّلَ لشيء مفيد (استحصي).
Alignment	تنسيق (توفيق)	أن يعالج النموذج حتى يُصدر مخرجا حسب تفضيلي (ما يوافق رغبتي).
Context window	نافذة السياق (حَدُّ السياق)	عدد محدد من الكُليمات يطلع عليها النموذج مرة واحدة.
Linear representation hypothesis	فرضية التمثيل الخطي	يُتفرض هنا أن الشبكات العصبية تمثل معانيَ مفيدة (مفاهيم لها معنًى) تُسمى ميِّزات
Superposition hypothesis	فرضية التراكب	يُفترض هنا صحة التمثيل الخطي ويُبنى عليه أن الشبكات العصبية تستعمل الاتجاهات شبه المتعامدة في فضاءات الأبعاد العديدة لتُمثِّلَ ميزات تزيد على عدد الأبعاد.
Category theory	نظرية الفئة	أحد أُطُر صوغ أسس الرياضيات الثلاثة (نظرية المجموعة والفئة والنوع)
Ecoder-Decoder	المشفِّر الفاك (مشفكك)	وهو الأصل والمشفكك نحتٌ
Deep learning	التعلم المُعَمَّق (المُشَعصب)	نحتنا جذر (مادَّة) ش ع ص ب اختصارا من الشبكة العصبية وسمَّيناها شعصبونة ويُمكن تصريف المادة الجديدة حسب الحال.
RAG	الاسترجاع المُعزز	استرداد المعلومات واستخدام النماذج لتجيب من المعلومة المستردة.
NLP	معالجة اللغة الطبيعية (المعالوية)	
LORA (low rank adaption)	التكايف الرتيب	

prompt engineering	هندسة الأوامر (الارشادات)	
python	اللغة البيثنية	عرَّبتُ الاسم ليسهل استخدامُه في المؤلفات ومنه جذر [ب ي ث ن].
Model	نموذج (منمذج)	أفضل استعمال وزن مُفَعلِل لأن فيه معنى الفعلية بدلًا من فعولل.
ML machine learning	تعلم الآلة	هذا هو المصطلح المشهور وأنا أستحسن "التعلم المُحَوسب" ليُتَخَصَّص المُتَعَلَّم بالذي كان في وسع الحاسوب.

المقالة الأولى في المُدَوَّنات وتمثيلها

فصل في جمع النصوص ومصادرها

بابُ أشهر المدوّنات

لا شك أن أهم ما في هذه النماذج هو هذا الحجم الهائل للنصوص المكتوبة المُستعملة في تدريبها (وسيأتي إن شاء الله منهج التدريب في الفصل القادم).

وأحسن ما نمثِّلُ به في مقام المدوَّنات العملاقة هي المدوّنات الكلزية وأكتفي بواحدةٍ منها اسمها "الشابكة الحسنة"[5] [6] عدد **كُليماتها** جاوز خمسَ عشرة **دَلفَ** كُليمة. انتُخبت من مدوّنات كلزية عديدة حوت مُعظمَ المعارفِ من كتبٍ وأوراقٍ علمية ومواقعَ منشورة على الشبكة انتُخبت بمساعدة النلضات، وأما المُدونات العربية فهي لا تضاهي حجما ولا تقارب فَحوًى هذه المدونة المهولة ولا ما دونها ولكنْ منها ما يَحسُن ذكرهُ واعتبارُه مُنطلقَ العملِ على المدونة العربية الخام (الصِّرفة):

1. مدونة 101 جَلْفَ كُليمة [7]: وهي مُدونة جديدة صدرت من فريق بحثي تونسي قاموا بانتقاء نصوصِها من مكانز النصوص العربية المجموعة من مزحوف الشابكة[6] مستعملين منهجية مُأتمتة (مُأيَّلة) لاختيار النصوص.
2. مدونة الثقافة السينية [8]: وهي مدونة عديدة اللغات (منها العربية) ومصدرها أيضا من مسحوبات الشابكة ويبلغ عدد كُليماتها العربية ما يُجاوز 70 جَلْفَ كُليمةٍ.
3. مدونة العربية 2022 وهي مدونة من تجميع شركة صينية حجمها 200 جيجابايت تم ترتيبها وتسويتها (حذف الحركات والمدود وغيرهما) ونُشرت على موقعهم وعدد كُليماتها حسب عدِّنا في المعمل 25 جَلْفَ كُليمة[7] [9].

★ وهناك مدونات التغريدات (ما نُشر وما أشير له) وآية (للمعيرة) نصوص كتبنا المسحوبة والويب العربي ٢٤ ومدونة C4

هذا هو أهمُّ ما نُشر على الشبكة لحدِّ كتابة هذه السطور وقد استعضنا عن ذكرِ نصوص الموسوكة ومدوّنات الكتب منفردةً[8] لأن غالبها مُضمَّن في المدوّنتين السابقتين، أو أنَّ حجمَ المُدوَّنة المتروكِ ذكرُها لا يدانيهم ولمن أحب الإطّلاع على غالب المنشور فعليه بموقع هَجِنْفِيْس[9]، ويحسن الإشارة أيضا إلى أن عددا ممن درَّبَ نماذجَ عربيةً قاموا بجمع مدونات أكبر من هذه ولكنَّها لم تُنشر كَمدوَّنة سدايا بـ380 جَلف كليمة ومدونة جيس بـ490 جلف كليمة وسيأتي الكلام عنهما وغيرهما في المقالة الثالثة.

[5] اسمها بالكلزية fineweb

[6] هذه تسمية الموقع Common crawler واختصارا تُكتب CC.

[7] العدُّ تمَّ بمُقسِّمنا Aranizer 64K وهو مرفوع على الشابكة https://github.com/riotu-lab/aranizer

[8] ألفت نظر الباحث إلى أن عددًا كبيرا من الكتب المهمة إلى الآن لم تُرَقمن وهذا من العجائب مثل قاموس المورد والمورد الأكبر وكذا عدد من المعاجم في مختلف المجالات كالمعرفة والفلسفة والكلام والعاميات والحضارة والقراءات وغيرها والموسوعات العامة العربية وكذا عدد من كتب الأدب كالأغاني والدر الفريد وبيت القصيد وغيرها كثيرٌ جدا من الكتب المصورة التي ينبغي نشرها بصورة نصية مفيدة.

[9] يُمكنُ الوصول إليه على هذا الرابط: https://huggingface.co/. فيه قسمٌ كامل للمُدوَّنات النصية والمرئية والصوتية وهو أعجوبةٌ.

باب طُرُق جَمع وتكوين المُدوّنةِ العربية

أشير في هذا الفصل إلى أهم الطرق المُستعملة في تجميع النصوص العربية وتكوينها وهي تنتظم في ثلاثةِ أصولٍ عامة:

1. أولها استنصاص الشابكة: وهي طريقة مُبَرمجة (عادة مُبَيثَنةٌ[10]) بحيث تُسحب النصوص من المواقع المنشورة، ويَجدُر الإشارة إلى أن عددًا كبيرا من هذه النصوص محفوظ ومجموعٌ خاصّة في زاحف الشبكة ولكن يعيبه الجودة والفحوى (أشير له في الفصل السابق).
2. ترجمة النصوص الكَلزيّة: وهي طريقَةٌ اسْتُعملت في عددٍ من النماذج العربية مثلَ جَيس [10]، أو عديدة اللغات مثل آية [11] حيثُ ترجمَ باحثو الورقتين نصوصا مُختلفة لتوسيع المدوّنة العربية المُستعملة في تدريب نموذجهم، وأنبِّهُ إلى أن هذه الطريقة ينبغي استعمالها وتطويرها لأن غالب العلوم مكتوبةٌ بالكلزية ولها فائدة أخرى وهي المساعدة في ترجمة العلوم بسرعةٍ وجودةٍ عالية.
3. استخراج نصوص الكتب: وهذه تُشبه الأصل الأول ولكنني أفردتها هنا لأن طُرُقها ومشاكلها مُختلفة، وأبيِّنُ هنا أمرين مهمّين الأوّل متعلقٌ بالكتب العربيّة المنشورة والتي غالبها بصيغة مبَدأفةٍ مُصوّرة غير رقمية لا يُستفاد منها إلا باستخراج نصوصها بتقنيات وخرزمات التعَرُّفِ البصري على المحارف وهذه أحدُ مشاكلها ضُعْفُ دِقَّتها في النصوص العربية وأما النصوص المكتوبة فقليل من له صلاحية الوصول إليها، وأما الأمر الثاني فهو مُتعلِّقٌ بالملفات العربية الرقمية أي الموجودة بصيغة رقميّة مُبَدأفة[11]، فهذه إشكالها أن تشفير النصوص العربية فيها غير مُنْرَجِع غالبًا[12]، فاستخدام المكتبات التي تَسحب النصوصَ من الملفَّات المُبَدأفة غير ناجع لأن أساس حفظها في هذه الملفات معطوب غير مُسترد.

وأيضا يجدر الإشارة لمصدر نصوص رابع ولكنّه قليلٌ مُنتَجُهُ، مُكلِفٌ بذلُهُ، وهي استكتاب الصوتيّات بتحويلها من مسموعٍ لمكتوب ومن أمثلته ما نشرته مؤسسة سدايا باسم صدى [12] وهو مجموعٍ صوتي جاوز طوله ستَّمائة ساعة مُسجَّلة من مُختلف البرامج من هيئة الإذاعة والتلفزيون السعودية.

فصل في طُرُقِ تمثيل النصوص

إن المتعامل مع الحاسوب يُدرك إدراكا جازمًا لا محيد عنه أنَّ كُلَّ ما يسري فيه ويُعالجُ أوّلُه وآخرُه أرقامٌ تتمَقحَلُ[13] في وحدة المعالجة المركزية وجاراتها في اللوحة الأم. ومن ذلك النصوص المكتوبة التي تَظهر أمامَ أعيُنِنَا في الشَّاشة وانطلاقا من ذلك فلا بُدَّ للنموذج اللغوي أن يَستقبل أرقاما تُمثِّلُ النصوصَ المرادَ تدريبُهُ عليها وينبثقُ من هذا الاقتضاء شرطُ إيجادِ خرزمةٍ لذلك ونشير إلى أشهرها وأكثرها استعمالا وهي خرزمة التزويج البَايتِيُّ (المزَبِّتة)، وكان أصل استعمالِها وتصميمِها لضغط النصوص[13]، ثمَّ ارْتُئيَ استعمالها في أول نموذج محول صدرَ عن جوجل في ورقتها المشهورة [14]. وهذه الخرزمة هي المُعتمدة في غالب النماذج اللغوية الكبيرة المنشورة إما مباشرةً أو مُتغيِّرةً بشكل بإطارِ "مُجْمِل" [15].

---

[10] البيثنة مُعرَّبة من اسم لغة البرمجة python فصككتُ منهُ جذرَ (ب-ي-ث-ن) فصار اسمها لغة بيثنة أو البيثنيَّة، وهناك من يستعمل برامج جاهزة للاستنصاص.

[11] البدأفة مُعَرَّبةٌ من الاختصار الكلزي pdf فصُغتُ منهُ جذرًا جديدا (ب-د-أ-ف) ثم من المصدر المؤنث بدأفةٌ.

[12] أي لا يُمكِنُ استرجاع المكتوب منها نصيا بعد حفظِهِ مُبدأفًا ومن أمثلة ذلك مكتبة pymupdf لا تستطيع استخراج النصوص العربية .

[13] المَقحَل هو transistor وهو الوحدة الأساسية في مُكوّنات الحاسوب.

باب شرح خرزمةِ التزويج البايتِيّ

تعتمد هذه الخرزمة على منهجية عدٍّ مختصرة تتكَّرر عددًا محدودًا من المرَّات (وهذا العدد غيرُ ذي حدٍّ مُعَيّن بل حسب التجارب والاختيار)، على التوالي، وقبل ذلك يتم تحويل النص المراد تدريبَ الخرزمة عليه إلى قائمة أرقامٍ تشفيرُ كُلّ حرفٍ منها يكون عادةً بترميز "أتف-8"[14].

مثال إن أردنا تدريب الخرزمة على الجملة التالية: "أنا طالب"

تَصِيْرُ قائمتها أرقامَها المُدخلة: [168 216 132 217 167 216 183 216 32 167 216 134 217 163 216]، ونَحفظ أيضا في البداية قوَيمسًا فيه الحروف المكوّنة للنص بدون تكرار كل مع رمزه الأتفيّ: {أ:216، ... } إلى آخره وهذه ستكون أصل قائمة المُفردات التي ستُزَادُ في الخرزمة.

ثم بعدها تبدأ عمليّة التزويج البَايِتِيِّ كالتالي:

1. يتم جَمعُ كُلِّ زوجين من الأرقام ويُحسب تكرار كل زوجٍ منها كالتالي:

{(216, 163): 1, (163, 217): 1, (217, 134): 1, (134, 216): 1, **(216, 167): 2**, (167, 32): 1, (32, 216): 1, (216, 183): 1, (183, 216): 1, (216, 167): 1, (167, 217): 1, (217, 132): 1, (132, 216): 1, (216, 168): 1}

2. يُستخرج أكثر الأزواج تكرارا من القائمة السابقة (216, 167) ويوضع في القائمة الأصلية برقم جديد (عادة يكون أكبر رقم في القائمة زائدًا واحدا):

[168 216 132 217 **(218)** 183 216 32 **(218)** 134 217 163 216]

3. تُكرَّر الخطوة الأولى عددا يُحدِّدُه الباحث (هذا العدد يُسمّى في الاصطلاح دمجات).

وهذا تعبير خرزمي عنها:

١: **تهيئة** القاموس بحيث يحتوي على كل حرف كعنصر
٢: **المدخلات**: النص
٣: تجزئة النص إلى حروف، لينتج قائمة من الرموز كرر
٤: **رِتَ**
٥: **الخطوة ١**: عدّ تكرار كل زوج من الرموز في النص
٦: **الخطوة ٢**: إيجاد الزوج الأكثر تكراراً
٧: **الخطوة ٣**: دمج الزوج الأكثر تكراراً في رمز جديد
٨: **الخطوة ٤**: تحديث قائمة الرموز في النص بالرمز الجديد المدمج
٩: **الخطوة ٥**: إضافة الرمز الجديد إلى القاموس
١٠: **نتِل** الوصول إلى العدد المطلوب من عمليات الدمج
١١: **المخرجات**: القاموس والنص المرمز

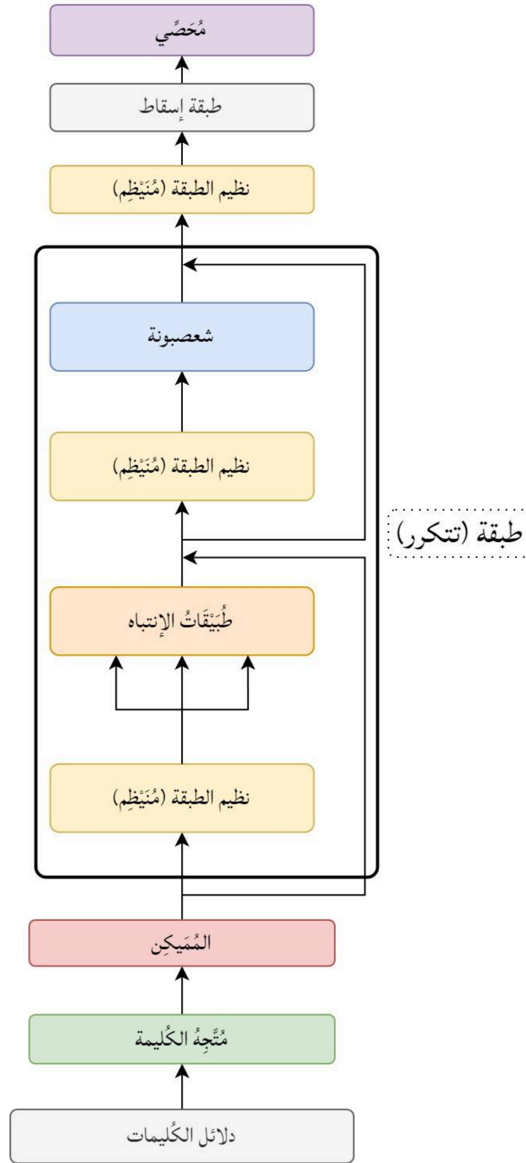
ولمن أراد رؤيتها مُبرمجة فليرجِعْ لما كتبناه في مقالة "الحَيْسُ من الجَيْسِ" [16] فقد فصَّلنا فيها عنها و أحلنا فيها على الشفرة المصدرية، وهنا بحث عن مسألة استعمال هذه المُقسمات مع العربية سلبا أو إيجابا (انظر بحث اليافعي ومقسم فنار)

---

[14] UTF-8

المقالة الثانية في شرح النموذج

في هذا الفصل نشرح الأصلَ الرياضيَّ وهيكلة النموذج المُحوّل فاكِّ التفشير والفرق بين مكوّناته واختلافها ذاكرًا أسماء أشهر النماذج اللغوية العربية، ولنبدأ بصورة عامّة للنموذج تُوَضِّحُ مكوِّناته الأساسية متبوعة بصورة لأهم معادلاته الرياضية لنحيل عليها عند الشرح إذ بعدها سنشرحه قطعةً قطعةً إن شاء الله وأُنَبِّه قبل ذلك إلى أنَّ هذه البُنية (الهيكلية) مختلفة قليلا عن هيكلية الفاكِّ من الورقة الأصلية [14] وتتبع ورقة متمم 2 :[17]



1. $X = [x_1, x_2, \ldots, x_T]$
2. $\text{Embeddings} = [E(x_1), E(x_2), \ldots, E(x_T)]$
   $\text{Positional Encodings} = [PE(1), PE(2), \ldots, PE(T)]$
   $Z = [E(x_1) + PE(1), E(x_2) + PE(2), \ldots, E(x_T) + PE(T)]$
3. $Q_i = Z^{(l-1)} W_i^Q$
   $K_i = Z^{(l-1)} W_i^K$
   $V_i = Z^{(l-1)} W_i^V$
   $\text{head}_i = \text{Attention}(Q_i, K_i, V_i)$
   $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$
   $\text{MHA}_{\text{self}}(Z^{(l-1)}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W_O$
4. $A_{\text{self}}^{(l)} = \text{LayerNorm}(Z^{(l-1)} + \text{MHA}_{\text{self}}(Z^{(l-1)}))$
5. $\text{FFN}(A_{\text{self}}^{(l)}) = W_2 \sigma(W_1 A_{\text{self}}^{(l)})$
6. $Z^{(l)} = \text{LayerNorm}(A_{\text{self}}^{(l)} + \text{FFN}(A_{\text{self}}^{(l)}))$
7. $\text{Final Norm} = \text{LayerNorm}(Z^{(L)})$
   $\text{Logits} = \text{Final Norm} W_{proj} + b_{proj}$

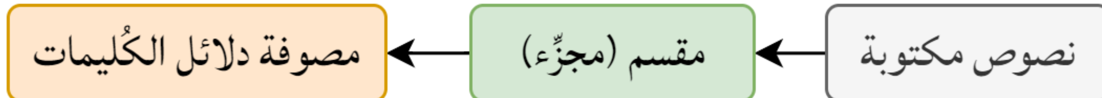
Dimensions:

- $Z^{(l-1)}$: $T \times d_{model}$
- W_i^Q, W_i^K, W_i^V: $d_{model} \times d_k$
- Q_i, K_i, V_i: $T \times d_k$
- head_i: $T \times d_v$
- W_O: $(h \cdot d_v) \times d_{model}$
- $\text{MHA}_{\text{self}}(Z^{(l-1)})$: $T \times d_{model}$
- $A_{\text{self}}^{(l)}$: $T \times d_{model}$
- W_1: $d_{model} \times d_{ff}$
- W_2: $d_{ff} \times d_{model}$
- $Z^{(l)}$: $T \times d_{model}$
- W_{proj}: $d_{model} \times V$
- Logits: $T \times V$

فصل في ما قبلَ الطبقة الجامعة

باب دلائل الكُليمات

في هذه المرحلة نحدِّدُ عددًا ثابتًا من الكُلَيمات T (هو مُخرَج المُقَسِّم) كما هو المعادلة الأولى (1) والتي سيتدرَّبُ عليها النموذج دُفعةً واحدةً، وهذا العدد ثابتٌ لا يتغيّر بعد اختياره ويُسَمَّى "نافذة السياق" أو "حَدّ السياق" لأنه يجعلُ الكُليمات تنفذ وتتطلع على بعضها لتتعلَّمَ كلٌّ منها ما يحتاجه من الأخرى، فإذا ما أتينا بعَيِّنةٍ (ولنفرض طولَ نافذتها 4 تسهيلًا)، فهي تكون قائمة فيها أربعة دلائل كلٌّ دليلٍ يُشير إلى كُليمتِهِ (وهذه الإشارة مهمة لأنّه عند الإنتهاء فإننا نُرجعُ الأدلة لهيئتها المكتوبة فينتُجُ نصٌّ مقروء). أَذَكِّرُ إلى أنَّ لدينا عددَ كُليمات محدودًا يُمثِّلُ النصَّ كله تم حفظه بعدَ تدريب خرزمة التزويج البتِّي.



باب مُتَّجِهِ الكُليمات

في هذه المرحلة يتم سحب مُتجهٍ مُمَيَّزٍ لكل دليلِ كُليمةٍ كما في المعادلة (2) (وبداية يكون هذا المُتَّجهُ عشوائيا يُتَعلَّم أثناء التدريب)، وهذه المُتَّجهات محفوظة في جدول كبيرٍ طوله عدد الكُليمات وعرضه عددٌ نختاره يكون طولَ (بُعدَ) متجه كل واحدة كما هو موضَّع في صورة الأبعاد التي على اليمين.

وأحب الإشارة إلى أن مَبحَث متجهات الكلمات عامَّة كان أحد الأسس التي قامت عليها نماذج اللغات بدءًا بورقة word2vec "كلمة لمتجه" [18] في 2013 ووصولا إلى ورقتنا ولمن أراد الاستزادة من هذا المبحث وعلاقته بالنلضات (النماذج اللغوية الضخمة) فليراجع مَبحثَ فضاء المعاني وقواعد حركة الدلالات من كتاب "كيف يَعمل متمم؟" [19].

باب المُمَيكِن (المموضع)

أبدأ بشرح معنى الكلمة فالمُمَيكِن مُفَيعِل من مكن وهي جذر المكان فهو الذي يضع الشيء في مكانه وناسب هذا لأن في هذه المرحلة[15] لأن فيها أيضا جدولَ متجهاتٍ ولكنَّه هذه المرة مُتعلِّق بتعليم النموذج مكان كُلِّ كُليمة إذ فَهْمُه اعتباطي لما يَدخُله فلا بُدَّ من تعليمه ذلك. ويَجدُرُ بالذكر أن طريقة تعليم النموذج مكان الكُليمات إحدى اثنتين: إما مُتَعلَّمةٌ أثناء التدريب وإما ثابتة بمعادلة يُحدِّدها الباحث (واسْتُعملت الثابتة في الورقة الأصلية للمحولات دالّة جا وجتا بسلوك دورانيٍّ [14]). واعلم أنْ ثَمَّ عددٌ كبيرٌ من طُرُقِ المَيكنة الثابتة منها ما استُعمل في نموذج جيس [10] وما استعمل في نموذج لاما (وهذه اشتهرت كثيرا وخاصة في مجتمع المصادر المفتوحة) [20] وغيرهما وأشهرها [21]: المُزيحُ الخطي ALIBI والتضمين الموضعي الدائري RoPE.

فصل الطبقة الجامعة

هذا الجزء من النموذج هو الجوهر والأساس إذ فيه أهمّ مُرَكَّب (قطعة) وهو طبقة الإنتباه (المتنابهة) وغالبُ عمليّة التعلّم الدلالي والمعنوي والسياقي فيها [22]، والطبقة الواحدة كما هو واضح في الشكل السابق مُكوّنة من ثلاث أجزاء: نظيم الطبقة، وطُبَيقات الإنتباه، وآخرها الشعصبونة (وهي اختصارٌ للشبكة العصبية) ولنشرحها جزءًا جزءًا

باب مُنظِم الطبقة

هي عمليّة تَنظِمُ صفوفَ المصفوفة لتَجعلها ذات متَوَسِّطٍ يُساوي واحدًا وتباينٍ يساوي الصفر، وهذا الإجراء مفيد في النماذجِ عميقة التركيب لأنها تحافظ على المُشتَقِّ العائد (المستعمل في خرزمة التدرج الاشتقاقي في الانتشار الخلفي) من أن يضمحلَّ فيستقر التعلم عدَدِيًّا وهي مُشكلة معروفة في هذا النوع من الشبكات [23] وهذه معادلة النظيم المُستعمل[16]:

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \epsilon}} * \gamma + \beta$$

وهي مشهورة الاستعمال في الدروس الإحصائية عموما فإنهم يُحِبُّون التوزيع الطبيعي (النِّظاميّ) الجَوْسِيّ، ويجدر أن نشير إلى أن في ورقة المحولات الأصليّة (وانظرها في المعادلة الرابعة (4) جعل المؤلفون هذه الطبقة قبل طبقة الاستنباه والشعصبونة وأما المُستعمل هنا فهو تصميم أصحاب نموذج متمم إذ هو المُعتمد في هذا النوع من نماذج المُحوّلات [17].

---

[15] وناسب كذلك المموضع لأنه من الموضع ومموقِع

[16] مُقتبسة من مكتبة بَتَيْرَش https://pytorch.org/docs/stable/generated/torch.nn.LayerNorm.html

وهذه هي أهمّ جزء من هذا النموذج إذ فيها عمليّة التواصل المتداخل بين كُليمات النصوص المُتَدَرَّبِ عليها، وكما هو موضح في المعادلات من القسم الثالث (3) فإن المتجهات (الناتجة عن جمع تضمينات الكُليمات ومتجهات الموضعة) تُجدى داخليًّا في ثلاث مصفوفات متساوية الأبعاد (انظر أبعادها في صورة الأبعاد اليُمنى) فينتج عنهُ ثلاثٌ جُدُدٌ همُ المستعملون في معادلة الإنتباه والتي تساعد في تواصل كل متجهٍ بمن حوله على ثلاث مراحل: الأولى مصفوفة Q وهذه تحاكي قولَك: "أنا أبحث عن معلومة معيَّنة" [انظر ترجمة query في المورد الحديث وغيره بمعنى الاستعلام و التساؤل والاستفهام صـ947]

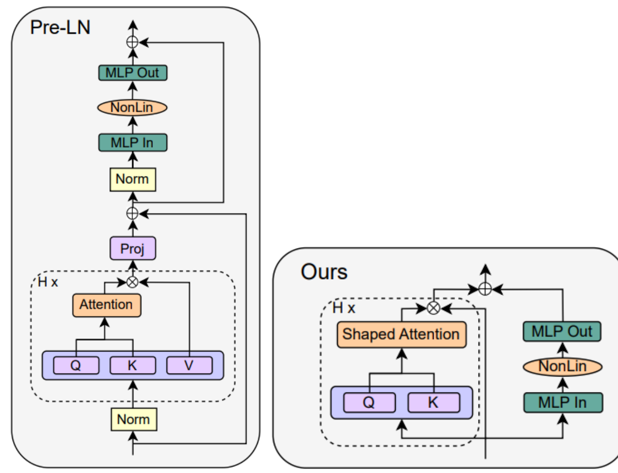
والثانية K والتي كأنها تقول: "أنا عندي المعلومة الفلانية" فهاتان عندما تُجدَيان تَعْرِفُ كلُّ كُليمة مع من تتفاعل وتتواصل، ثم بعد ذلك يُحَصَّوْنَ (وسيأتي شرح المُحَصّي) ويُنَظَّمون (يُنَسَّبون) مقسومين على جذر بُعد d_k (انظره في قسم الأبعاد)، وأخيرا يُجدى (يُضرب) الناتج بمصفوفة القيَم V وهي التي تُقَيِّم أهم النتائج وتَسْتَخْلِصُ منها.[17] وأشير باختصار إلى أنَّ هذه المعادلة تُسمَّى "الإنتباه الداخلي" لأنها تعمل بين نفس الكُليمات تفريقًا بينها وبين "الإنتباه المتشابك" المُستعمل في نماذج الترجمة ذات بُنية المشَفِّر الفاكّ. وأشير أيضا إلى أن الجمعَ في الطُبيقات هو جمع تصغير الطبقة وهذا أسلوب في تكثير المحسوبات المتآزية وهو أن تُقسم طبقة الإنتباه الواحدة إلى عدد مُعَيَّن من الطُبيقات تُتَعَلَّمُ وتَنحسبُ كُلُّها في وقت واحد ثم بعد ذلك تُضمُّ مُخرجاتها في متجه واحد وهو المقصود من معادلة MHA وهي اختصار لـ multi head attention أي انتباه عديد الرؤوس واختصارا طُبيقة انتباه وكلمة concat المراد بها الضم أو الإلصاق لعدد من المتجهات وجعلها واحدًا.

ولعل سائلا يسأل إن كانت الكُليمات تتفاعل مع بعضها فلا بُدَّ أن أوائلها تتواصل مع المُستقبَلِيَّات (أواخرها) أفليسَ ذلك غشٌّ إذ يُمكنهم الإطلاع على ما سيأتي بعدهم في نموذج التوليد؟ نقول هذا صحيح ولذلك في النماذج التوليدية فمعادلة الإنتباه (3) يكون فيها زيادة تمنع اطلاع الكُليمة على غيرِ ما قبلها وهي كالتالي:

$$\text{Attention}(Q, K, V, M) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V$$

فهذه المصفوفة الجديدة M هي مصفوفة مربعة أبعادها مثل أبعاد المصفوفة الناتجة تضع مكان كُلِّ ما يُراد حجبه (تقنيعه) قيمةً لانهائيةَ سالبةً فينتج عن ذلك بعد التحصية مصفوفةٌ كُل ما نريد حجبه عن الكُليمة فيها يصير صفرًا. ويحسن أن نشير إلى أن بعض الباحثين حاول دراسة هيكلية المحولات وما يُمكن الاستعاضة عنهُ من مكوِّناته وتأثير ذلك على أداءه حتى أنه جعل الانتباه بمصفوفتين فقط [24] وهاتان الصورتان فالتي على اليسار هي بالهيكلية (البُنية) المشهورة والتي على اليمين هي المُهَذَّبة بما استُعيضَ عنهُ فلاحظ فروقهما.

[17] **ملاحظة:** الأصل في نماذج الشعصبونات أنها خَفيَّة (مُلغزة) العلَّة غير واضحٍ فيها مَا يُتعلَّم والتشبيهات السابقة تقريبية اجتهادية كلٌّ حسبَ فهْمِه وأريَحُ للمرء أن يفهمَ حسابيَّتها وإن أحب الاستزادة فعليه بمجال "التعلم المُفَسَّر" وأُسميه المُعَلَّل وسنشير له في فقرة مُختصر تأتي إن شاء الله



باب طبقة الشبكات العصبية

وهذه من عتيق الطبقات بل هي الأصل والأساس لكل بُنى (هيكليَّات) نماذج التعلم العميقة بأنواعها وأشكالها. وتعريفها: تقريبٌ دالّيٌّ يتعلَّم نمط (علاقة) المُدخلات (المُعطيات) بالمُخرجات مُستعملا التراكيب الخطيَّة[18] ودالَّة التفعيل. ومكوِّنات الشبكة العصبية ثلاثة: مُدخل ذو بُعدٍ محدد و طبقة وسيطة (مخفيَّة) فيها أوزان ترتبط بالمُدخلات خطيًّا ثم دالة تفعيل تُضفي سلوكًا غير خطيٍّ على مجموع ما قبلها وتُصاغ اختصارا كما في المعادلة الخامسة بتعبيرات مصفوفاتية (5) ولا بُدَّ أن نشير إلى أن الدالة **σ** تعبير عامٌّ يُمكن استعمال أي دالة تفعيل فيه، ومن أشهرها استعمالا دالَّة GELU وكذا عائلة GLU ففيها عدة أشكال [21]. وأختصر هنا الكلام عن **طبقة الإسقاط** إذ هي شعصبونةٌ بلا دالَّة تفعيل. وهذه المعادلة هي الصورة العامة للشبكة العصبية بدون متغيِّر الانحياز b. ولمن شاء اطلاعا على تاريخ الشعصبونات فلينظر [77]

$$h_i = \sigma(W_i h_{i-1})$$

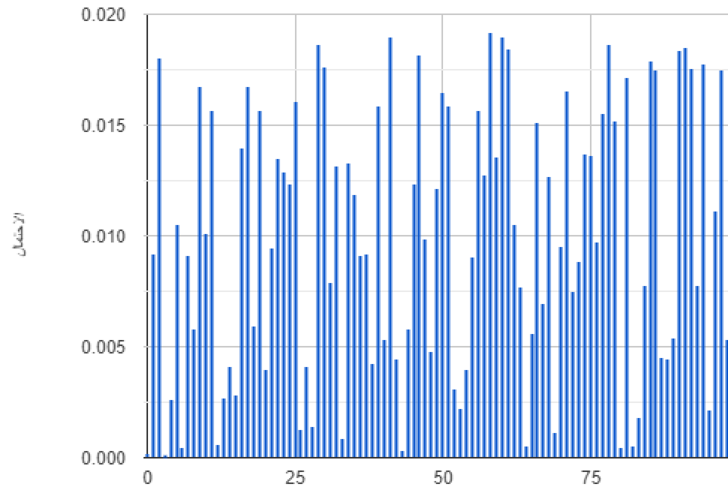
فصل فيما بعد الجامعة وأُخَر

باب المُحَصِّي

بعدَ المرور على آخر طبقة من طبقات "الطبقة الجامعة" -وهي التي تكررت عددا من المرات- نصل إلى آخرِ ثلاث طبقات من النموذج وهي المُنظِم والمُسقِط والمُحَصِّي. قبل الوصول إلى المُحَصِّي (وسأعَرِّفُه قريبا) يكونُ مُخرجُ طبقة الإسقاط متجهات بعدد المُدخلات طول كلٍّ منها عدد الكُليمات وانظر المعادلة السابعة (7) واسمها مُحُض (أو خوالص) وهذه أرقام حقيقية تَدخُلُ بَعدها على دالة التحصية softmax والتي تحوِّلها إلى توزيع إحتمالي (انظر الصورة) نختار منه الكُليمة الجديدة إما بسحب أعلى احتمال أو سحْبٍ احصائي من التوزيع.

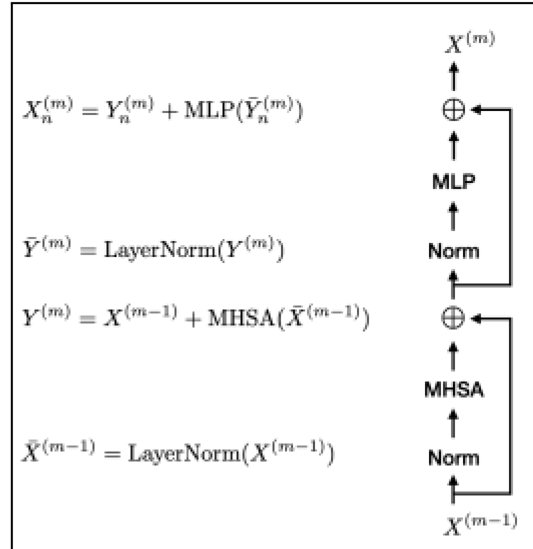
$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

[18] التراكيب الخطيَّة هنا هي معادلات من الدرجة الأولى ax+b



باب جريان المُدخَلات

يلاحظ الناظر في هيكلية المحول الفاكّ أن هناك أسهمًا تخرج من المُدخَل قبلَ أن يلجَ إلى جزءٍ معيَّن في موضعين من الطبقة الجامعة وهذا يُسمَّى "بالجريان" Residual وهو أن يجمع المُدخَلُ صِرفًا (دون تغيير) مع مُخرج الطبقة التي تليه وهذه الطريقة تساعد على استقرار المُشتق العائد في الانتشار الخلفي وكذا تُعِيْنُ النموذج على أن يتدارك ما يتعلَّمه إن أخطأ في الطبقة المُستعملة فيستعمل المُدخل صرفًا في تحسين التدريب وهذه رسمة ومعادلات تساعد على التصور [25]:



فقرة مختصرة عن أسباب تعلم هذه النماذج

في هذه الفقرة نشير إلى مبحثٍ يتنامى جِرمُه ويتبازغ نجمُه في الفترة الماضية حولَ تفسير وتعليل تعلُّمِ النماذج اللغوية الكبيرة وسأختصر القول فيه من جهتين: الأولى فيما يتعلق بمركبات المحول وربط الأُسُس المُتَعَلَّمة من علم اللسانيات بكل جزءٍ منها ففي ورقة "آليات المعرفة في النماذج اللغوية الضخمة" [22] عُرِضَت وصيغَت معادلات رياضية لتحليل المعرفة داخل النماذج اللغوية واستخلصوا خمسَ فرضيات عن توزيع المعرفة في النماذج اللغوية أولها: أن المعرفة تُشَفَّر (تُرَمَّز) في مناطقَ مقاسيّة[19] والثانية: أن المعرفة تُمَثَّل بالوصلات والمقصود بالوصلات هنا هو جريان المدخل في قنوات النموذج والثالثة: النماذج اللغوية الكبيرة تستعمل عناصر محددة عندَ استيعاب المعرفة وتطبيقها

---

[19] مقاسيّة هنا تعني Modular

والرابعة: النماذج اللغوية الكبيرة تُشَكِّل (تُكَوِّن) المعرفة من خلال الاستقراء. والخامسة: التناقض والتكامل يتواجدان في تطور المعرفة المتحركة في النماذج اللغوية. وكذا ذكروا مواضع ما يُتعَلَّم مخصوصا من الحفظ والإبداع والفهم المعنوي والنحوي وغير ذلك.
وأما الجهة الثانية: فهي في تمثيل المعرفة رقميا ففي "دراسة مسحية عن انشراح (قابليَّتهُ للشرح) النماذج اللغوية" [26] قسموا الطرق حسب مراحل التدريب وبينوا مواضع التفسير أيضا في مُدخلات الأوامر وكذا في التقييم وفي "أوليات عمل جوهر نماذج المحولات اللغوية" [27] قسموا التفسير إلى سلوك محليٍّ وتشفير المعلومات وأخيرا اكتشاف السلوك الداخلي وطبعًا كل واحدةٍ منها تصاغ في معادلات رياضية ولن أقدر على التفصيل فيها خوفَ الإطالة. وأما في أبحاث شركة أنثربيك فإن عندهم قسمًا كاملا في الشركة لهذه المسألة ومن ذلك في بحثهم "إطار رياضي لدوائر المحولات" [28] قاموا بتدريب نموذج صغير جدا ومحاولة دراسة أرقامه بأساليب جبرية مختلفة وأما في بحثهم "توسيع المعاني المتعددة: باستخراج ميِّزات مُنفسرة من كلود 3 سونيت" [29] فقد استطاع الباحثون اسْتِخْرَاج مَيِّزاتٍ واضحة ذاتِ معنًى مجرد يفهمه الإنسان مثل مفهوم علوم الدماغ وأخطاء البرمجة كاتجاهات في فضاءات التفعيل وكل ذلك تحت ظن صحة فرضيّة "**التمثيل الخطيّ ثم فرضية التراكب**". وممن درس هذه النماذج دراسةً رائعة خلط فيها التنظير بالتطبيق سلسلة أوراق باسم "فيزياء النماذج اللغوية الضخمة". [78]
وأختم هذه الفقرة بلمحةٍ سريعة إلى علمٍ برز في الفترة السابقة هدفُه تنظير الشبكات العصبية وبناء منهج رياضي متينِ الأساس تُفهم فيه الشعصبونات بكل هيكليَّاتها تحت أصول رياضية موحَّدة فمن ذلك ورقة باسم "نظرية الشبكة العصبية" [30,53] وكذا ورقة عن "هندسة التعلم العميق" [31] وهذه حاول مؤلفوها تنظير قاعدةٍ بنظرية الفئات وتطبيق فكْرَ إرلانجن على مجال التعلم العميق وأيضا ففي ورقة "فئوية التعلم العميق ونظرية جبرية لكل الهياكل" [32] هناك نفس الفكرة. وما سبق هو إشارة مقتضبة جدا لعل الله ييسر أن ينفتح بها بابٌ لتطبيق ذلك وتعريبه حسب حاجة المجال البحثي العربي وممن درس بعض ذلك في النماذج العربية عام 2022 [76].

تنبيهات

الأول: ما سبقَ من الشرح كانَ تقريبًا لغالب المفاهيم بصوغها صياغةً مصفوفية والواقع أن غالب بلْ كُلُّ النماذج الحقيقية يتم استعمال المرصوصات في تدريبها نظرًا لجودة وفعاليّة كروت الشاشة في حسابِ عمليَّاتها وسيزيد ذلك في المعادلات السابقة بُعدًا جديدا تُرصُّ فيه أمثلة أكثر يتدرَّبُ عليها النموذج تُسمَّى "دُفعة" وهي (أي الدفعة) أمثلة معدودة يتدرَّبُ عليها النموذج مرة (دُفعةً) واحدةً [79].
الثاني: أن تدريب هذه النماذج مُكلفٌ حوسبيًّا جدا سواءً في طور التدريب الأول وحتى في طور الإستعمال [33] ولذلك دأبت الأبحاث على التلاعب والتغيير لمركباته تارةً أو تحميله تارةً أخرى طلبًا للسرعة وتوفيرا للوقت والتكلفة ومن أحب الاطلاع فعليه بهذه الورقة [34] وهي استقصاء لأساليب زيادة جودة وفعالية النماذج من الجهة التدريب والاستعمال ومن جهة البيانات وتحسينها وكذا فلينظر أوراق التكاليف الكهربية [80].
والتنبيه الأخير والثالث وهو مما أستخلصه من بحثي في المجال أن أهمَّ أمرٍ في تدريب النماذج اللغوية (وغيرها من النماذج العميقة) هو معرفة استعمال المكاتب البرمجية المُخصصة لتدريب النماذج وأشهرها pytorch التيرشيَّة فإن المشاكل أثناء التدريب لا تُعدُّ ولا تكاد تُحصى فينبغي لمن أراد الولوج لمجال تدريب النماذج مباشرةً أن يُقَعِّدَ أصولًا صلبة في فهم أسس التدريب بدءًا من بناء النماذج وطُرقُ كتابتها وتمثيلها وتسريع التدريب وتحسين الأرقام العوائم ومناهج التآزي ولْيُكثِرِ النظر في المكتبات المنشورة. وللفائدة فإن أحدَ أهمِّ أسباب طفرة انتشار تدريب نماذج المحولات الفاكَّة ورقة من ستنفورد باسم "الانتباه السريع" Flash Attention قامت بتغيير طريقة حساب دالة التنبُّه لتتناسب كُليًّا مع كروت الشاشة من شركة نفيديا [35].

باب في تقييم النماذج

س: هذا الباب لم أكتبه بعدُ لأن الورقة لم يُقصد استيعابها لكل أوجه النماذج اللغوية وعليك في توسيع الورقة النظر في كل المصادر التي تجمعها في نوشن وغيره مما تكتبه (لعلها تتحول إلى كتاب في الباب).

تقييم النماذج في المسائل العربية الثقافية بحيث نسأله من علوم الإسلام مع العربية أسئلة دقيقة في الشرح والاستحضار تدل على ضحالته لغويا

A Survey on Evaluation of Large Language Models

A Survey of Useful LLM Evaluation https://arxiv.org/pdf/2406.00936

فصل في أشهر النماذج العربية

أذكر هنا في هذا القسم أشهر النماذج العربية التي نُشرت في الفترة الماضية مما اتَّبع هيكلية فاكِّ التشفير ولن نطيل فيها كثيرا بل نكتفي بالاختيار المختصر والتسمية إذ "مرصد المجمع" فيه فصل مُفصَّل عن النماذج العربية احصاءًا لها ولتطبيقاتها فلا داعيَ للإطالة هنا عنها.

1. سلسلة نماذج AraGPT: وهذا أول نموذج عربية [36] يتبع هيكلة متمم طوَّرَه وسام أنطون ومن معه عام 2020 على ما يقارب 80 جياتًا[20] من البيانات النصية المنوَّعة وصدر منه أربعة أحجام على التوالي: (والـM تعني بَلفًا في عدِّ المتغيرات): 135M, 370M, 792M, 1460M.
2. سلسلة نماذج AceGPT وهي نماذج عربية أصلها معيرة نماذج لاما 2 إذ تم تدريب هذه النماذج على مجموعات نصوص عربية جاوزت 19 دَلْفَ كُليمة عربية لتعوِّدها على النصوص العربية وطبيعتها وصدر منها نموذجان: 30B, 13B (حيثُ B هنا تعني جَلفًا -مليار-) وقريبا صدرَ إصدارٍ ثانٍ لهذه النماذج [37].
3. سلسلة نماذج Jais: ولعل هذه أشهر السلاسل العربية والتي صدرت عن جامعة محمد بن زايد في الإمارات بالتعاون مع شركة أنظمة كريبراس سنة 2023 أول نموذج كان حجم 13B وتدرب على ما يقارب 110 جَلْفَ كُليمة، ثم عام 2024 صدرت عائلة كاملة جديدة جاوزت العشرين نموذجا متنوعة الأحجام من 500 بَلفِ متغير وصولا إلى 70 جَلْفًا [10, 38].
4. سلسلة نماذج علام: وهذه نماذج طورتها شركة سدايا في المملكة العربية السعودية فيها عائلتان عائلة معيرت نماذج لاما 2 وعائلة دُربت مُستَأْنَفَةً[21] وفيها عددٌ كبير جدا من الكُليمات ما بين أصلي ومُترجم وكذا استعملوا أسلوب خلط اللغتين العربية والكلزية وحسب تجاربهم فإنه أدى إلى نتائج أفضل. [39]
5. سلسلة آية Aya: وهو من تطوير شركة كوهير وهي نماذج مفتوحة المصدر في نسختين الأولى 8B والثانية 35B وبياناتها كذلك مفتوحة المصدر [40]، وهي مدربة على 23 لغةً منها العربية والكلزية والصينية وغيرها.
6. سلسلة بَلوم bloom: وهذه تُعتبر من أوائل النماذج عديدة اللغات مفتوحة المصدر أكبر واحدٍ فيها حجمه 176B وأصغرها: 560M طورتها جهة Big Science عام 2023 [41]. ومن أحب أن ينصدم فلينظر في الورقة أول صفحتين ونصف يُلفها كُلها في أسماء العاملين على النماذج.
7. نموذج ArabianGPT: هذه النماذج صدرت من معملنا معمل الروبوتات وإنترنت الأشياء في جامعة الأمير: صدر منه إلى الآن أربعة نسخ، بنفس أحجام نماذج GPT 2 وما زلنا نعمل على تطويرها وتحسينها بعون الله [42].

[20] أيضا من تعريبات الوحدات [2] وهي لـ Gigabyte GB فوحدة byte هي بيات ثم يُستبدل أول حرفٍ بالقيمة ج

[21] أذكر أن الاستئناف يعني البدء من الصفر وليس الإكمال.

مقالة معيرة النماذج وتطبيقاتها

في هذه المقالة نشير إلى طُرُق الاستفادة من النماذج اللغوية إذ إنَّ الملاحَظَ من الفصل الماضي أن النموذج توليديٌّ صِرفٌ لن يفهم سؤالا ولن يعِيَ أمرًا وعليه فلا بُدَّ من أسلوب أو طريقة لتوجيهه أو قُل لترويضه على أساليب مُعيَّنة يُستفاد بها منه، وأيضا فإن النموذج لما يُنتهى من تدريبه فإن معلوماته تثبت ولن تتغير من تغيرها في الواقع وهذا أيضا شقَّ مجالا جديدا لحل هذه المشكلة.

فصل في المَعيرة المُهيكَلة

بعدَ تدريب النموذج يَنتجُ (يصير) عندنا كاتبٌ مولِّدٌ يُعطَى أول الكلام فيكتب من عنده ما أحبَّ دون أسلوب معين فهو احتوى (حفظ) المعلومات التي تدَرَّب عليها ولكنَّه لا يفهم الأمر (والأمر يشمل السؤال وغيره من التطبيقات اللغوية كالترجمة والتصنيف) فاستَلزُم أن نعوِّده على أسلوب كتابة المُدَردشات (مُدخَل يحتوي أمرا يستدعي إجابةً) مُستعملين المعيرة المُهيكلة [43] على النهج التالي:

- يتم جمع عدد كبير من الأمثلة (عادة مئات الآلاف وبعضهم أوصلها لعشرة ملايين) ذات بِنية محددة (نمط ثابت) تُحدَّد بكُليمات خاصة كالمثال التالي: [<مستخدم> ... أمر معين ... <\مستخدم> <نموذج> ... جواب ... <\نموذج>] وهذا النمط كل جهةٍ مدربة للنموذج تصوغه بطريقتها ما بين مُطَوِّلٍ ومُقَصِّرٍ وانتبه أن هذه الرموز تُعتبر كُلُّ واحدةٍ منها كُليمةً خاصة يعرفها المقسم.
- يُدرب النموذج كتدريبه العادي على هذه الأمثلة حتى يتعوَّد على هذا الأسلوب ويصيرَ فاهمًا لمبدأ السؤال والجواب.

ومن المدونات المهيكلة (أو المرتبة) لمعيرة النماذج العربية (وهي قليلة وتحتاج بحثًا مطولا نظرا لركاكة كثيرٍ منها بسبب ترجمتها آليًّا من اللغة الكلزية والذي أدى إلى إنتاج نماذج تكتب بحروف عربية معانيَ أعجمية):

1. مدونة Bacterian X: وهذه صدرت من جامعة محمد بن زايد لعدد من اللغات من العربية في 67 ألف أمر وجوابه [44].
2. مدونة آية من شركة كوهير وهي أكبر هذه المذكورات إذ حجم النصوص العربية التي فيها يقارب 25 جيجا بايت وفيها ملايين الأمثلة العربية المتنوعة ما بين مجموع ومترجم [45].
3. مدونة الـ 500 ألف وهذه صدرت من مركز بحثي اسمه ClusterlabAI وتحتوي على 500 ألف مثال ما بين مجموع ومولَّد (مُصَنَّع) في مجالات عديدة مثل التصنيف والأسئلة والشرح والتلخيص وغيرها [46].
4. مدونتنا: وهي مجموعة أسئلة قُمنا بجمعها في المعمل من عدد من المدونات للنماذج ذات السياق القصير 1024 كُليمة [47].
5. مدونة ألباكا Alpaca: وهي مدونة أصدرتها ستنفورد في ورقة بحثية لهم قاموا بكتابتها باستعمال نموذج متمم 3.5 عام 2023 [48] وترجمها عدد من الناس للعربية باستعمال نماذج الذكاء الاصطناعي وعددها خمسون ألف عينة وهي موجودة على huggingface.

ونستطرد هنا لأمرٍ مهمٍّ في تدريب متمم الثاني وكذا الثالث [17,49] أضافوا مرحلة ثالثة (بعد التدريب الأولي والمعيرة المهيكلة) ألا وهي تهيئة (تحذية) النموذج على أجوبة مخصوصة (بأسلوب معين) حتى لا يجيب على أي سؤال يُسأل عنه (وهذا له اعتبارات أخلاقية وأمنية وشخصية) باستعمال ما يُسمَّى "التعلم المعزز من الإفادات البشرية" وهذه المرحلة فيها تفاصيل كثيرة وتفرعات عديدة انبثق عنها طرقٌ مختلفة ما بين مُحَسِّن ومجدد وأشير إلى أشهرها بعناوينها ومن أحب فعليه الإطلاع على دراسة مسحية صدرت قريبا جمعتها كلها وشرحتها وقد بلغت 33 أسلوبَ تنسيق (تحذية) [50]:

- التعلم المعزز من إفادات الذكاء الاصطناعي reinforcement learning from AI feedback RLAIF
- استمثال التفضيل المباشر Direct Preference Optimization DPO
- استمثال كهمان تفرسكي Kahneman-Tversky Optimization KTO

فصلُ أهم تطبيقات النماذج اللغوية الضخمة

أشهر تطبيق للنماذج اللغوية الضخمة (ومعلوم المراد في هذه الورقة كما سبق في المقدمة هم فاكُّو التشفير) بعد معيرتها هو الدردشة (أو الاستجابة للأوامر) وأشهر هذه التطبيقات مما يدعم العربية:

- مدردش شركة أُبِن آي OpenAI المشهور باسم Chat GPT وله عدة نسخ آخرها عند كتابة هذه الأوراق GPT-4 Mini ما بين مدفوع ومجاني.[22]
- مدردش شركة جوجل واسمه Gemini وله كذلك عدة نسخ ما بين مجاني ومدفوع.[23]
- مدردشا شركة كوهير Cohere الأول اسمه آية 23 Aya وهما مفتوحا المصدر ولهما نسختان الأولى ذات 8 مليارات متغير والثانية ذات 23 جَلْفَ متغير. ويُعتبران امتدادا لنموذجهم المُسمَّى كُمَاندر Command R المنشور قبلهما وكل هذه مجانيٌّ استعمالها.[24]
- مدردش شركة كَلود Claude المسمى سُنِتْ 3.5 وهذا نُشر قبل فترة قريبة من الشهر السادس عام 2024 وهو مجاني الإستعمال.[25]
- مدردش شركة مسترال آي Mistral AI المسمى "الكبير الثاني" Large 2 ولعل هذا أحدث ما في هذه الورقة من النماذج الداعمة للعربية.[26]

و أحب أن أشير إلى أشهر نموذج مفتوح المصدر واسمه "لاما" LLAMA تنتجه شَركة ميْتا Meta وقد صدر إلى وقت كتابة هذه السطور ثلاثة نسخٍ [20, 51] منه ولكنه مقتصرٌ على اللغة الكلزية إلا آخر نسخة ففيها 6 لغات كلها لاتينية وفيه دعمٌ للعربية ولعله غير مقصود.
وكذا يحسن الإشارة إلى بزوغٍ نجمِ علمٍ جديد يتعلَّقُ بهندسة الأوامر (أو كيف تستخرج من النموذج أحسن إجابة) وقد صدرت دراسة مسحية جمعت غالب أساليبه وشرحتها وبلغت 39 أسلوبًا خلاصتها في تخصيص الأمر قدر الإمكان وتنسيقه حوَت 29 مهمة لغوية [55]، وأيضا هناك فرعٌ بحثي يتعلَّق بإعطاء النموذج البيانات الجديدة -وأشرنا لهذه المشكلة أول الفصل- ليجيب منها على الأسئلة بدلًا مما تعلمه من معلومات قديمة وهذا المجال صار بينِيًّا بين علم استرداد المعلومات وعلم النماذج اللغوية وقد صدرت دراسة مسحية جمعت غالب الطرق والفروق بينها وربط

[22] https://openai.com/chatgpt/

[23] gemini.google.com

[24] coral.cohere.com

[25] claude.ai

[26] https://mistral.ai/news/mistral-large-2407/

كل طريقة بالمهمة اللغوية المناسبة [56] وخصَّ بعضُ الباحثين العرب دراستها عن العربية فقط [57]. وأختم أيضا بالتنبيه على مسألة مهمة جدا وهي فهم الثقافة العربية في النصوص المُستعملة في التدريب فقد يُؤدي عدم الاهتمام إلى توليد النموذج لنصوص تخالف تماما ثقافتنا وديننا وقد أشار بعد الباحثين لذلك ما بينَ دارسٍ للظاهرة وبين مقيمٍ للنماذج على المهام اللغوية العربية [58:63]

باب في التطبيقات الفرعية

كل الشركات السابقة عدا ميتا تقدم نماذجها بواجهة للمستخدمين مع واجهة برمجة تطبيقات API ليستخدمها المبرمج في مُنتجاته سواءً كانت مواقع أو تطبيقات ويجدر الإشارة إلى أن عددًا كبيرا من المبرمجين يستعملون هذه النماذج لتُهَيكِلَ النصوص غير المرتبة بصيغٍ مشهورة ك"جسَن" json و"يمَل" YAML وغيرها، وهذه الفكرة أدَّت إلى تطوير عددٍ مهول من التطبيقات والبرامج التي تستعمل هذه النماذج. ولا شك أن استعمال النماذج أثر على غالب المجالات الأساسية من هندسة البرمجيات بجميع مراحلها [64]، والطب والصحة [65] والقانون [66] والماليات [67] والرياضيات [68] (وآخرها من دربته جوجل [69] على إثبات أسئلة أولمبياد الرياضيات) وغيرها وهذه دراسة مسحية فيها أهم تطبيقات وأهمها الدردشة في مجالات متنوعة وتكوين البيانات المُصطنعة وكذا أهم تحدِّياتها وهي في أصلين إما في سوء المُخرج وطريقة إصلاحه أو في نفس المشكلة المراد حلها [70,71]، وهي كثيرة منها ما يتعلق بالجانب العلمي عند الناس ومنها ما يتعلق بالتكاليف وغيرها وأحب أن أشير إلى أمر مهم ينبغي دراسته وهو تأثير هذه النماذج على التحصيل الدراسي وما يُفَرَّعُ على ذلك من غش وتغيير في النظام التعليمي وهياكله وأسئلته ومناهجه [72]. ولعل قارئا يريد دراسةً موسوعية حول معالجة اللغة العربية فمن كان كذلك فعليه بهذه الدراسة المرأوية[27] وللأسف أنها في عام 2021 ولعل الله ييسر أن تصدر نُسخة أحدث منها [73].

خاتمة وتوصيات

عرضنا (بل إن شئت قُل ضغطنا وحشرنا) في هذه الورقة عُلالَةً مُختصرة وإشارات مُقتصرة حول نماذج المحولات التوليدية العربية وما يتعلق بها من مُدوَّنات وخرزمات. فقد شرحنا مركبات النموذج ومن ثَمَّ أشهر النماذج العربية ثم ختمت بتطبيقات النماذج اللغوية. وأختم بأمر يجول بخاطري حول علاقة العربية بهذا المجال من مُنطلقين: الأول حول ضرورة الإهتمام بها ابتداءًا قبل التسارع إلى إدخالها في النماذج اللغوية أو غيرها من التطبيقات الحاسوبية فإنه لا يُتصور أن تكتب الأبحاث وتُنشر المقالات بلغة غيرها والمُتكَلَّمُ عنه هي. والثاني حول المجالات المُهتَمِّ بها وأولويَّتُها في خدمة العربية، فإن الناظر يُدرك بلا شك أن غالب العلوم الآن منشورة بالكلزية ولا بُدَّ أن تُعرَّب وتُعدَّل المناهج كذلك لذلك فأظن أنه ينبغي الإهتمام بدءًا بالترجمة وتكثير بحوثها وتطويرها حتى نقدر على تقديم هذه العلوم بأصولها وأُسُسُها بلسان عربي ثم ننطلق بعد ذلك لضغط هذه المعارف وصهرها في النماذج اللغوية. وأخيرا أختم بنصيحة استللتها من كلام تشومسكي لمَّا سُئل عن هذه النماذج فقال "أنها لم تُقَدِّم شيئا في المجال العلمي وإنما هي أدوات مفيدة هندسيا وتطبيقيًّا" [74] أقدمِّها للباحث لألا يترك ما بين يديه من بحوث نظرية ومسائلَ علمية ويَهرَع لهذه المُندَلقات يَكتُب حولها بل فليَلزَم كراسته وقلمَه وبَحثه فإنما هي زوبعةٌ في فنجان وجعجعة بلا إطحان إذ:

(من راقب الناس لم يظفر بحاجته... وفاز بالطيبات الفاتك اللهج)[28]

و (يا باريَ القوس بريا ليس يُحسنُه ... لا تظلم القوس أعط القوس باريها)[29]

---

[27] المرأوية هي ترجمة panorama وانظر إن شئت المصدر [75]

[28] هذا البيت لبشار بن بُرد

[29] ص319 - كتاب شرح الشواهد الشعرية في أمات الكتب النحوية المكتبة الشاملة وفيه أنه لم يُعرف قائله.

المراجع والمصادر[30]

[1] شينون ي., & بن سعد م. ا. (2022). مشكلة المصطلحات اللسانية في اللغة العربية ـ دراسة في مصطلح الفونيم The Problem of Linguistic Terms in the Arabic Language - A study of the term Phoneme /// (Phoneme). القارئ للدراسات الأدبية والنقدية واللغوية, 5(1), 80-99. https://www.asjp.cerist.dz/en/article/183991

[2] كتاب تسمانية وحدات القياس في اللغة العربية، خضير شعبان 1441 من ديوان اللغة العربية diwanalarabia.com

[3] معجم الرياضيات، مجمع اللغة بدمشق، الدكتور موفق دعبول ومن معه، 2018

[4] كتاب مختار الصحاح صـ18 - أز ا - المكتبة الشاملة. وفيه أن نبه وقال "ولا تقل وازاه"

[5] معجم مصطلحات التعلم الآلي والتعلم العميق وعلم البيانات د.علاء طعيمة.

[6] FineWeb: decanting the web for the finest text data at scale https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1

[7] Aloui, M., Chouikhi, H., Chaabane, G., Kchaou, H., & Dhaouadi, C. (2024). 101 Billion Arabic Words Dataset. arXiv preprint arXiv:2405.01590. https://huggingface.co/datasets/ClusterlabAi/101_billion_arabic_words_dataset

[8] Nguyen, T., Van Nguyen, C., Lai, V. D., Man, H., Ngo, N. T., Dernoncourt, F., ... & Nguyen, T. H. (2023). Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. arXiv preprint arXiv:2309.09400. https://huggingface.co/datasets/uonlp/CulturaX

[9] https://data.baai.ac.cn/details/ArabicText-2022

[10] Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., ... & Xing, E. (2023). Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. arXiv preprint arXiv:2308.16149.

[11] Üstün, A., Aryabumi, V., Yong, Z. X., Ko, W. Y., D'souza, D., Onilude, G., ... & Hooker, S. (2024). Aya model: An instruction fine tuned open-access multilingual language model. arXiv preprint arXiv:2402.07827.

[12] https://www.kaggle.com/datasets/sdaiancai/sada2022

[13] Philip Gage. 1994. A new algorithm for data compression. The C Users Journal, 12(2):23–38.

[14] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[15] Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226.

[16] الحيس من الجيس مقالةٌ شارحة لنموذج جيس 2023 سري السباعي: https://caramel.la/serrysibaee/YxM4Fnlps/alhys-mn-aljys

[17] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[18] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.

[30] اعتمدت على أسلوب APA في نقل المراجع فيما وجدته وإلا فكتبت اسم الكتاب والمؤلف وتاريخ الطبعة وقد قال الغرناطي [ومن يجد من خطأٍ أو من زلل ... أذنت في إصلاحه لمن فعل] ولا ينبغي للكاتب أن يمدح مكتوبه ولكنّه فضل الله فلينظر الناظر فيه بعين العدل والاستفادة والحمد لله وحده على فضله.

[19] Wolfram, S. (2023). What Is ChatGPT Doing:... and Why Does It Work?. Wolfram Media. Section: "Meaning Space and Semantic Laws of Motion" https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

وهذه الورقة أيضا درست شيئا من ذلك بما يتعلق بالمفاهيم المجردة التصنيفية

Park, K., Choe, Y. J., Jiang, Y., & Veitch, V. (2024). The Geometry of Categorical and Hierarchical Concepts in Large Language Models. arXiv preprint arXiv:2406.01506.

[20] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[21] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

[22] Wang, M., Yao, Y., Xu, Z., Qiao, S., Deng, S., Wang, P., ... & Zhang, N. (2024). Knowledge mechanisms in large language models: A survey and perspective. arXiv preprint arXiv:2407.15017.

[23] Ba, J. L. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

[24] He, B., & Hofmann, T. (2023). Simplifying transformer blocks. arXiv preprint arXiv:2311.01906.

[25] Turner, R. E. (2023). An introduction to transformers. arXiv preprint arXiv:2304.10557.

[26] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology, 15(2), 1-38.

[27] Ferrando, J., Sarti, G., Bisazza, A., & Costa-jussà, M. R. (2024). A primer on the inner workings of transformer-based language models. arXiv preprint arXiv:2405.00208.

[28] A Mathematical Framework for Transformer Circuits, Nelson Elhage and others 2021 https://transformer-circuits.pub/2021/framework/index.html

[29] Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, Adly Templeton and others 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

[30] Petersen, P. C. (2020). Neural network theory. University of Vienna, 535.

[31] Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478.

[32] Gavranović, B., Lessard, P., Dudzik, A. J., von Glehn, T., Araújo, J. G. M., & Veličković, P. Position: Categorical Deep Learning is an Algebraic Theory of All Architectures. In the Forty-first International Conference on Machine Learning.

[33] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). An empirical analysis of compute-optimal large language model training. Advances in Neural Information Processing Systems, 35, 30016-30030.

[34] Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., Qu, Z., ... & Zhang, M. (2023). Efficient large language models: A survey. arXiv preprint arXiv:2312.03863, 1.

[35] Dao, T., Fu, D., Ermon, S., Rudra, A., & Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems, 35, 16344-16359.

Dao, T. (2023). Flash Attention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691.

[36] Antoun, W., Baly, F., & Hajj, H. (2020). AraGPT2: Pre-trained transformer for Arabic language generation. arXiv preprint arXiv:2012.15520.

[37] Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Song, D., ... & Xu, J. (2023). AceGPT, Localizing Large Language Models in Arabic. arXiv preprint arXiv:2309.12053.

Zhu, J., Huang, H., Lin, Z., Liang, J., Tang, Z., Almubarak, K., ... & Xu, J. Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion.

[38] https://huggingface.co/inceptionai/jais-adapted-70b

[39] Bari, M. S., Alnumay, Y., Alzahrani, N. A., Alotaibi, N. M., Alyahya, H. A., AlRashed, S., ... & Khan, H. (2024). ALLaM: Large Language Models for Arabic and English. arXiv preprint arXiv:2407.15390.

[40] Aryabumi, V., Dang, J., Talupuru, D., Dash, S., Cairuz, D., Lin, H., ... & Hooker, S. (2024). Aya 23: Open weight releases to further multilingual progress. arXiv preprint arXiv:2405.15032.

[41] Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., ... & Bari, M. S. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

[42] Koubaa, Anis, Adel Ammar, Lahouari Ghouti, Omer Necar, and Serry Sibaee. "ArabianGPT: Native Arabic GPT-based Large Language Model." (2024).

https://huggingface.co/riotu-lab

[43] هذه دراسة مسحية عن أساليب المعيرة المهيكلة

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., ... & Wang, G. (2023). Instruction tuning for large language models: A survey. arXiv preprint arXiv:2308.10792.

[44] https://huggingface.co/datasets/MBZUAI/Bactrian-X

[45] Singh, S., Vargus, F., Dsouza, D., Karlsson, B. F., Mahendiran, A., Ko, W. Y., ... & Hooker, S. (2024). Aya dataset: An open-access collection for multilingual instruction tuning. arXiv preprint arXiv:2402.06619.

[46] https://huggingface.co/datasets/ClusterlabAi/InstAr-500k → paper: GemmAr: Enhancing LLMs Through Arabic Instruction-Tuning

[47] https://huggingface.co/datasets/riotu-lab/ArabicQA_2.1M

[48] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., ... & Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html, 3(6), 7. https://crfm.stanford.edu/2023/03/13/alpaca.html

[49] Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[50] Wang, Z., Bi, B., Pentyala, S. K., Ramnath, K., Chaudhuri, S., Mehrotra, S., ... & Asur, S. (2024). A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. arXiv preprint arXiv:2407.16216.

[51] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md (llama 3)

[52] Mahboub, A., Za'ter, M. E., Alfrou, B., Estaitia, Y., Jaljuli, A., & Hakouz, A. (2024). Evaluation of Semantic Search and its Role in Retrieved-Augmented-Generation (RAG) for Arabic Language. arXiv preprint arXiv:2403.18350.

[53] Otto, S. E., Zolman, N., Kutz, J. N., & Brunton, S. L. (2023). A unified framework to enforce, discover, and promote symmetry in machine learning. arXiv preprint arXiv:2311.00212.

[54] Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2024). Datasets for large language models: A comprehensive survey. arXiv preprint arXiv:2402.18041.

[55] Vatsal, S., & Dubey, H. (2024). A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. arXiv preprint arXiv:2407.12994.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927.

[56] Ding, Y., Fan, W., Ning, L., Wang, S., Li, H., Yin, D., ... & Li, Q. (2024). A survey on rag meets llms: Towards retrieval-augmented large language models. arXiv preprint arXiv:2405.06211.

[57] Mahboub, A., Za'ter, M. E., Alfrou, B., Estaitia, Y., Jaljuli, A., & Hakouz, A. (2024). Evaluation of Semantic Search and its Role in Retrieved-Augmented-Generation (RAG) for Arabic Language. arXiv preprint arXiv:2403.18350.

[58] Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2023). Having beer after prayer? measuring cultural bias in large language models. arXiv preprint arXiv:2305.14456.

[59] Alghamdi, E. A., Masoud, R. I., Alnuhait, D., Alomairi, A. Y., Ashraf, A., & Zaytoon, M. (2024). AraTrust: An Evaluation of Trustworthiness for LLMs in Arabic. arXiv preprint arXiv:2403.09017.

[60] Alyafeai, Z., Almubarak, K., Ashraf, A., Alnuhait, D., Alshahrani, S., Abdulrahman, G. A., ... & Al-Shaibani, M. S. (2024). CIDAR: Culturally Relevant Instruction Dataset For Arabic. arXiv preprint arXiv:2402.03177.

[61] Almazrouei, E., Cojocaru, R., Baldo, M., Malartic, Q., Alobeidli, H., Mazzotta, D., ... & Noune, B. (2023, December). AlGhafa Evaluation Benchmark for Arabic Language Models. In Proceedings of ArabicNLP 2023 (pp. 244-275). https://gitlab.com/tiiuae/alghafa

[62] Khondaker, M. T. I., Waheed, A., Nagoudi, E. M. B., & Abdul-Mageed, M. (2023). Gptaraeval: A comprehensive evaluation of chatgpt on Arabic nlp. arXiv preprint arXiv:2305.14976.

[63] Alyafeai, Z., Alshaibani, M. S., AlKhamissi, B., Luqman, H., Alareqi, E., & Fadel, A. (2023). Taqyim: Evaluating Arabic nlp tasks using chatgpt models. arXiv preprint arXiv:2306.16322.LAraBench: Benchmarking Arabic AI with Large Language Models Elmadany, A., Nagoudi, E. M. B., & Abdul-Mageed, M. (2022). ORCA: A challenging benchmark for Arabic language understanding. arXiv preprint arXiv:2212.10758.

[64] Zhang, Z., Chen, C., Liu, B., Liao, C., Gong, Z., Yu, H., ... & Wang, R. (2023). Unifying the perspectives of nlp and software engineering: A survey on language models for code. arXiv preprint arXiv:2311.07989.

[65] Zhou, H., Gu, B., Zou, X., Li, Y., Chen, S. S., Zhou, P., ... & Liu, F. (2023). A survey of large language models in medicine: Progress, application, and challenge. arXiv preprint arXiv:2311.05112.

[66] Lai, J., Gan, W., Wu, J., Qi, Z., & Yu, P. S. (2023). Large language models in law: A survey. arXiv preprint arXiv:2312.03718.

[67] Li, Y., Wang, S., Ding, H., & Chen, H. (2023, November). Large language models in finance: A survey. In Proceedings of the fourth ACM international conference on AI in finance (pp. 374-382).

[68] Liu, W., Hu, H., Zhou, J., Ding, Y., Li, J., Zeng, J., ... & He, L. (2023). Mathematical language models: A survey. arXiv preprint arXiv:2312.07622.

[69] AlphaGeometry: An Olympiad-level AI system for geometry
Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. Nature, 625(7995), 476-482.

[70] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. arXiv preprint arXiv:2307.10169.

[71] Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Gao, C., ... & Zhao, Y. (2024). Trustllm: Trustworthiness in large language models. arXiv preprint arXiv:2401.05561.

[72] Pudasaini, S., Miralles-Pechuán, L., Lillis, D., & Salvador, M. L. (2024). Survey on Plagiarism Detection in Large Language Models: The Impact of ChatGPT and Gemini on Academic Integrity. arXiv preprint arXiv:2407.13105.

[73] Darwish, K., Habash, N., Abbas, M., Al-Khalifa, H., Al-Natsheh, H. T., Bouamor, H., … & Mubarak, H. (2021). A panoramic survey of natural language processing in the Arab world. Communications of the ACM, 64(4), 72-81.

[74] لقاء بعنوان "الشبح في الآلة" ghost in the machine على المَيتب youtube مع تشومسكي قناة Machine learning street talk
https://www.youtube.com/watch?v=axuGfh4UR9Q

[75] Abd al-Raḥīm, Y ، موسوعة العامية السورية: دراسة لغوية نقدية في التفصيح والتأصيل والمولد والدخيل. صـ 410

[76] Abdelali, A., Durrani, N., Dalvi, F., & Sajjad, H. (2022). Post-hoc analysis of Arabic transformer models. arXiv preprint arXiv:2210.09990.

[77] Bishop, M. J. (2015). History and philosophy of neural networks.
Wang, H., & Raj, B. (2017). On the origin of deep learning. arXiv preprint arXiv:1702.07800.

[78] https://physics.allen-zhu.com/home (هذا الموقع يحوي كل أوراقهم مرتبة مع مرئيات تزيدها شرحًا)

[79] Panagakis, Y., Kossaifi, J., Chrysos, G. G., Oldfield, J., Nicolaou, M. A., Anandkumar, A., & Zafeiriou, S. (2021). Tensor methods in computer vision and deep learning. Proceedings of the IEEE, 109(5), 863-890.

[80] Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2023). Estimating the carbon footprint of bloom, a 176b parameter language model. Journal of Machine Learning Research, 24(253), 1-15.
Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., & Jiang, L. (2023). Llmcarbon: Modeling the end-to-end carbon footprint of large language models. arXiv preprint arXiv:2309.14393.